ABSTRACT OF THE INVENTION

A method and system for mining a document containing dirty text. Dirty text is removed or replaced and the document is processed using a variety of text mining techniques. In one embodiment, dirty text removal and replacement occurs in two stages. In the first stage, a general cleaning occurs on all documents without regard to what domain they belong to or the mining task to be performed. In the second stage, document cleaning is more specific to the anomalies of the domain and the mining task to be performed . In the third stage, the document is processed using a variety of data mining techniques according to the mining task. In one embodiment, the present invention scores and ranks sentences in a document according to their relevance, extracts the highest ranked sentences, and presents a summary. The present invention allows users to leverage existing domain knowledge and can be customized according the domain and task requirements.